*D252–D257* *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 27 November 2012*
*doi:10.1093/nar/gks1168*

# miRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting

**Sooyoung Cho[1], Insu Jang[2], Yukyung Jun[1], Suhyeon Yoon[1], Minjeong Ko[1], Yeajee Kwon[1], Ikjung Choi[1], Hyeshik Chang[3], Daeun Ryu[1], Byungwook Lee[2], V. Narry Kim[3], Wankyu Kim[1,*] and Sanghyuk Lee[1,2,*]**

[1]Ewha Research Center for Systems Biology (ERCSB), Ewha Womans University, Seoul 120-750, [2]Korean Bioinformation Center (KOBIC), KRIBB, Daejeon 305-806 and [3]School of Biological Sciences, Seoul National University, Seoul, 151-742, Korea

Received September 15, 2012; Revised October 27, 2012; Accepted October 28, 2012

## ABSTRACT

**Biogenesis and molecular function are two key subjects in the field of microRNA (miRNA) research. Deep sequencing has become the principal technique in cataloging of miRNA repertoire and generating expression profiles in an unbiased manner. Here, we describe the miRGator v3.0 update (http://mirgator.kobic.re.kr) that compiled the deep sequencing miRNA data available in public and implemented several novel tools to facilitate exploration of massive data. The miR-seq browser supports users to examine short read alignment with the secondary structure and read count information available in concurrent windows. Features such as sequence editing, sorting, ordering, import and export of user data would be of great utility for studying iso-miRs, miRNA editing and modifications. miRNA–target relation is essential for understanding miRNA function. Coexpression analysis of miRNA and target mRNAs, based on miRNA-seq and RNA-seq data from the same sample, is visualized in the heatmap and network views where users can investigate the inverse correlation of gene expression and target relations, compiled from various databases of predicted and validated targets. By keeping datasets and analytic tools up-to-date, miRGator should continue to serve as an integrated resource for biogenesis and functional investigation of miRNAs.**

## INTRODUCTION

Over the past 2 years, the number of known microRNAs (miRNAs) in human has almost tripled (1). The catalog of miRNA information is usually deposited in databases such as miRBase (1) and PMRD (2). In miRNEST (3), novel miRNA candidates are predicted from expressed sequence tag (EST) sequences in various animals, plants and viruses. The miRNAs of related sequences are grouped as RNA family as in Rfam (4).

Regarding miRNA targets, validated targets are still sparse but are available at miRecords (5), Tarbase (6) and miRTarBase (7). Many target prediction methods were developed including TargetScan (8), microRNA.org (9), miRBase (1), PITA (10), PicTar (11), miRDB (12) and their combinations (13). These programs usually suffer from a large number of false positives. Other tools that provide analytics functions based on miRNA and mRNA expression profiles include HOCTAR (14) and miRFANS (15).

The biology of miRNAs is turning out to be much more complex than initially thought, where a single miRNA may have multiple isoforms (iso-miRs) and often undergo modifications such as 3′-nucleotide addition (16). Comprehensive profiling of such miRNA variants is necessary to understand the function of miRNAs in the context of various human diseases and other perturbations. Deep sequencing technique is rapidly replacing the hybridization-based methods due to its ability to catalog and quantify miRNAs (and their variants) in an unbiased and accurate manner. Accordingly, several web tools and databases, including deepBase (17), miRTools (18), miRanalyzer (19) and miRDeepFinder (20), were developed to analyze the deep sequencing data.

*To whom correspondence should be addressed. Tel: +82 232772888; Fax: +82 232773760; Email: sanghyuk@kribb.re.kr
Correspondence may also be addressed to Wankyu Kim. Tel: +82 232774132; Fax: +82 232773760; Email: wkim@ewha.ac.kr

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Even though deep sequencing has become the main driving force in uncovering novel miRNAs and expression changes, we still lack a comprehensive and integrated database of miRNA sequencing, expression profiling and targeting information, implemented with proper tools. Here, we introduce the miRGator v3.0 that consolidated an extensive datasets of deep sequencing studies. The user interface is fully renovated with a dedicated miRNA-seq browser and two novel viewers that enable users to examine the miRNA–target relationships with expression correlation information readily accessible. We describe the main characteristics of the updated system in the following sections.

## SYSTEM OVERVIEW

The schematic overview of miRGator v3.0 is shown in Figure 1. We have included deep sequencing data available in public, which have become the principal resource for information on miRNA diversity and expression. The datasets were manually curated into ontology-based disease and tissue categories. We have compiled 73 studies with 4665 samples into 38 disease and 71 anatomic categories.

Major features, summarized in Figure 1, include (i) miR-seq browser, which allows users to examine short read alignment for identifying iso-miRs and differential expression in multiple samples; (ii) expression profiles in various organs, tissues and diseases, based on deep sequencing data; (iii) novel representation of miRNA–target relations in correlation heat-maps and network views of gene expression and (iv) gene set analysis for functional annotation of miRNA-associated genes.

## DATASETS AND PROCESSING OF SEQUENCING DATA

We have collected 73 deep sequencing datasets on human samples from Gene Expression Omnibus (GEO) (21), Short Read Archive (SRA) (22) and The Cancer Genome Atlas (TCGA) archives (23). GEO and SRA included 54 studies of miRNA and mRNA sequencing (716 samples and 4.1 billion short reads). Additionally, we added the expression profiles of miRNAs and mRNAs in cancer samples from the TCGA archive (19 studies, 3949 samples in 17 cancer types). TCGA data are particularly useful in investigating the inverse expression correlation of miRNA and target mRNAs in various types of cancer. Note that the TCGA level 3 data include the processed output only, not the raw sequence data. All GEO/SRA experiments and TCGA data were manually annotated into tissue and disease types using the controlled vocabulary of eVOC (24) and MeSH (25), respectively. Table 1 shows the summary of datasets included in this update.

The miRNA deep sequencing data were aligned to the reference human genome (hg19) using the Bowtie program (version 0.12.7) (26) after trimming adaptor sequences by Cutadapt (version 1.1) (27) obtained from the original paper or manufacturer platform. Up to two mismatches were allowed in the alignment process to identify iso-miRs or miRNA modifications. Short reads mapped onto the known miRNA loci from miRBase v18 (1) or ncRNA region from Ensembl (release 67) (28) were classified as miRNA or ncRNA reads, respectively. This procedure yielded 1856 known miRNAs and 6424 ncRNAs. Remaining reads were used to predict novel

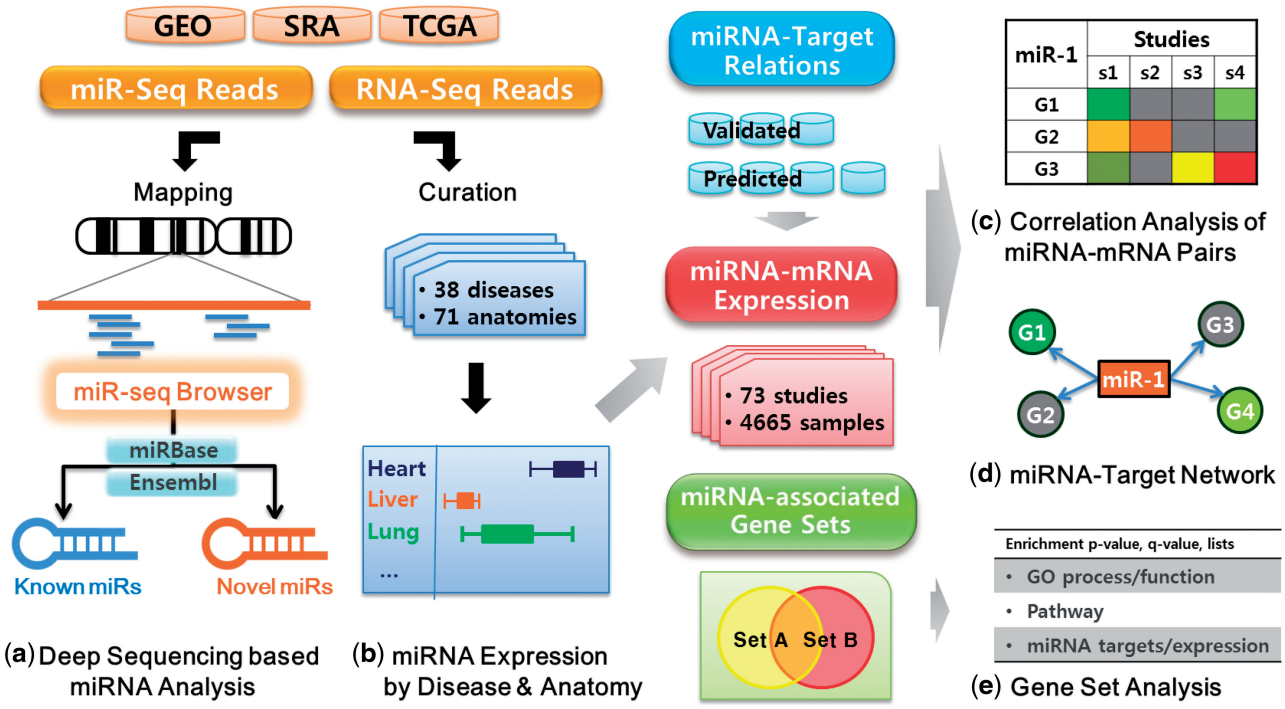


**Figure 1.** System overview of miRGator v3.0.

**Table 1.** Statistics for deep sequencing data and curation result

| | | GEO | SRA | TCGA | Total |
|---|---|---|---|---|---|
| Curation | No. of studies | 44 | 10 | 19 | 73 |
| | No. of samples | 660 | 56 | 3949 | 4665 |
| | No. of anatomies | 54 | 15 | 18 | 71 |
| | No. of diseases | 26 | 4 | 17 | 38 |
| Mapping | No. of total reads | 3 651 203 657 | 545 986 295 | – | 4 197 189 952 |
| | No. of trimmed reads | 2 704 297 513 | 147 800 838 | – | 2 852 098 351 |
| | No. of mapped reads | 2 129 934 409 | 392 826 996 | – | 2 522 761 405 |
| | No. of mapped reads to miRNAs region | 1 663 515 565 | 286 992 242 | – | 1 950 507 807 |
| | No. of mapped reads to ncRNAs region | 108 819 368 | 20 060 074 | – | 128 879 442 |
| | No. of mapped reads to genomic region | 191 686 502 | 22 757 497 | – | 214 443 999 |
| Processing result | No. of pre-miRNAs | 1521 | 1429 | 747 | 1522 |
| | No. of mature miRNAs | 1843 | 1661 | 934 | 1856 |
| | No. of other ncRNAs | 6421 | 6286 | – | 6424 |
| | No. of predicted pre-miRNAs | 286 | 69 | – | 304 |
| | No. of predicted mature miRNAs | 475 | 94 | – | 508 |

miRNAs using the mirDeep2 software (29). Using the estimated true-positive probability of 95% and randfold *P*-value of 0.05, we obtained 508 mature and 304 pre-miRNA candidates. Further details of the analysis pipeline and program options are available in the online documentation.

For quantification of miRNA abundance, we used the quantile normalization method for read numbers within each miRNA locus. Differentially expressed miRNAs (DEmiRs) between tumor and normal tissues were obtained by edgeR program (version 2.6.10) (30) after converting the normalized number into the nearest integer value.

RNA-seq data were aligned to the human genome (hg19) by the TopHat program (version 2.0.0) (31) after removing adaptor sequences and critical examination of quality controls. Cufflinks (version 1.3.0) (32) was used to quantify the mRNA abundance.

## miR-seq BROWSER

miR-seq browser was specifically designed to examine the sequence alignment and normalized read counts with the secondary structure information in an intuitive and interactive fashion. Short reads related to iso-miRs and miRNA editing can be readily identified with the corresponding expression values (read counts) in multiple samples. This feature can be of significant value for scientists studying biological roles of iso-miRs and miRNA editing.

Figure 2 shows the screen shot of miR-seq browser. The secondary structure, obtained from Vienna RNA package (33), is displayed on the top panel and also indicated as different shades in the alignment window. Selecting each nucleotide in the secondary structure highlights corresponding nucleotide in the sequence alignment panel. Mismatch sequences are indicated in red color. Users may add, delete or edit read sequences. The read count table can be used to explore the variable expression of iso-miRs and differential miRNA processing. Expression level is also reflected as the background color of each cell in this table. We have further implemented many user-friendly features such as zoom-in/out, reordering of reads (drag & drop), sorting by expression level and save/restore support of configuration. It is also possible to upload the user sequences in the BAM file format. Detailed instructions for using miR-seq browser are available in the online help page.

## miRNA, TARGET mRNA AND EXPRESSION CORRELATION

Inferring molecular functions of miRNAs is a non-trivial process due to the uncertainty in relationships between miRNA and target mRNAs. Only small portions of target mRNAs are known for a limited number of miRNAs, and typical programs tend to yield too many false positives. We have compiled a variety of miRNA–mRNA relationships and integrated them with the expression correlations to help users identify reliable targets readily.

Validated miRNA target genes were obtained from miRecords (version 3), mirTarBase (version 2.5) and Tarbase (version 5). Predicted target relationships were collected from Microcosm Targets (version 5) (34), miRDB (version 4), miRNA.org (August 2010), PITA (version 6), PicTar (May 2004) and TargetScan (version 6.2). In total, miRGator v3.0 includes 4745 validated and 6 218 792 predicted target relations, nearly doubled from the previous version.

Expression correlation is useful information to discern between direct and indirect targets. Inversely correlated expression of miRNA and putative target mRNAs is a strong evidence for genuine relations. We calculated the correlation coefficient using the deep sequencing data of mRNA-seq and miRNA-seq from the same sample. We used the Spearman rank correlation which is robust to different normalization methods between mRNA-seq and miRNA-seq data.

Target relation and expression correlation are visually represented in two formats as shown in Figure 3. The heat-map view shows the expression correlation between miRNA and target mRNAs within each dataset. The

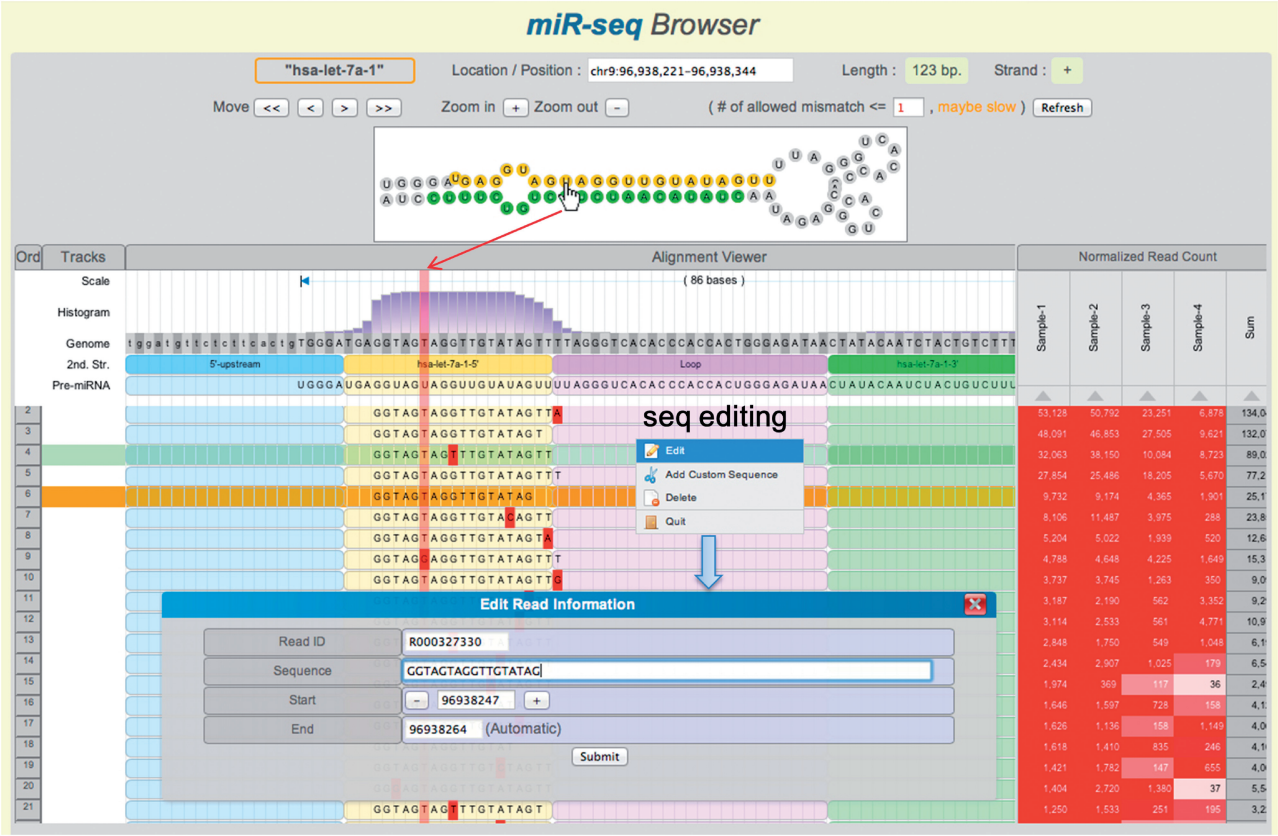


**Figure 2.** Main features of the miR-seq Browser. At the top panel, the hairpin structure of miRNA precursor is shown. The aligned short reads are shown together with secondary structure and read depth information in the track. By mouseover (hand icon) on a nucleotide, the corresponding columns are highlighted in vertical pink shadow. The reads can be sorted by the read count of each sample or the total sum on the right panel. Note that several read sequences show 3′-end modifications. Histogram shows the read depth at each position. Mismatched nucleotides are highlighted in red. Sequence editor window is opened by right-click.

scatter plot of miRNA and mRNA expression can be displayed by clicking each cell to examine sample-dependent variation of gene expression. The source of target information is also indicated to help users identify consensus targets, which are more likely to be genuine targets (13). All information on target relationship and expression correlation is downloadable in Excel format to allow more elaborate analysis for users.

Network view shows the target relationship in the graph visualization format. Users may select the validated or predicted target relations, study ID of source data and samples. Gene expression levels or the fold changes, if applicable, are shown as the node color. Network view illustrates the target relations and expression correlations in more intuitive manner, but limited to display the expression in a single study or sample. It should be noted that the miRNA–mRNA relation can be queried either by the miRNA name or by the gene name, which is a useful feature to investigate any synergistic effect in miRNA or gene function (35).

## GENE SET ANALYSIS

Gene Set Analysis (GSA) is commonly used in interpreting a list of genes from high-throughput experiments such as microarray and mass spectrometry. The GSA tool of miRGator v3.0 enables the user to compare a list of genes against *a priori* defined gene sets such as KEGG pathway, Gene Ontology, the validated/predicted miRNA target DBs and inversely coexpressed gene sets as described in the previous section. The statistical significance is calculated as *P*-value by hypergeometric test, which is corrected for multiple tests using Bonferroni method.

## USER INTERFACE

The miRGator v3.0 website incorporates various user-friendly features. Most menus are self-evident except the miR-seq browser for which detailed instructions are available in the help page. Basic search can be performed for miRNA, disease and anatomy names. The search window suggests plausible keywords and supports the auto-complete mode.

Search output for miRNA query consists of (i) basic information including GeneRIF information, (ii) relevant studies, (iii) samples in the selected study where the link to miR-seq browser is available and (iv) miRNA expression profiles in disease, tissue and organ categories. Anatomy or disease queries output relevant studies and the DEmiRs from each study. Search in the 'miR-target & Expression menu' can be performed for miRNA or gene of interest,

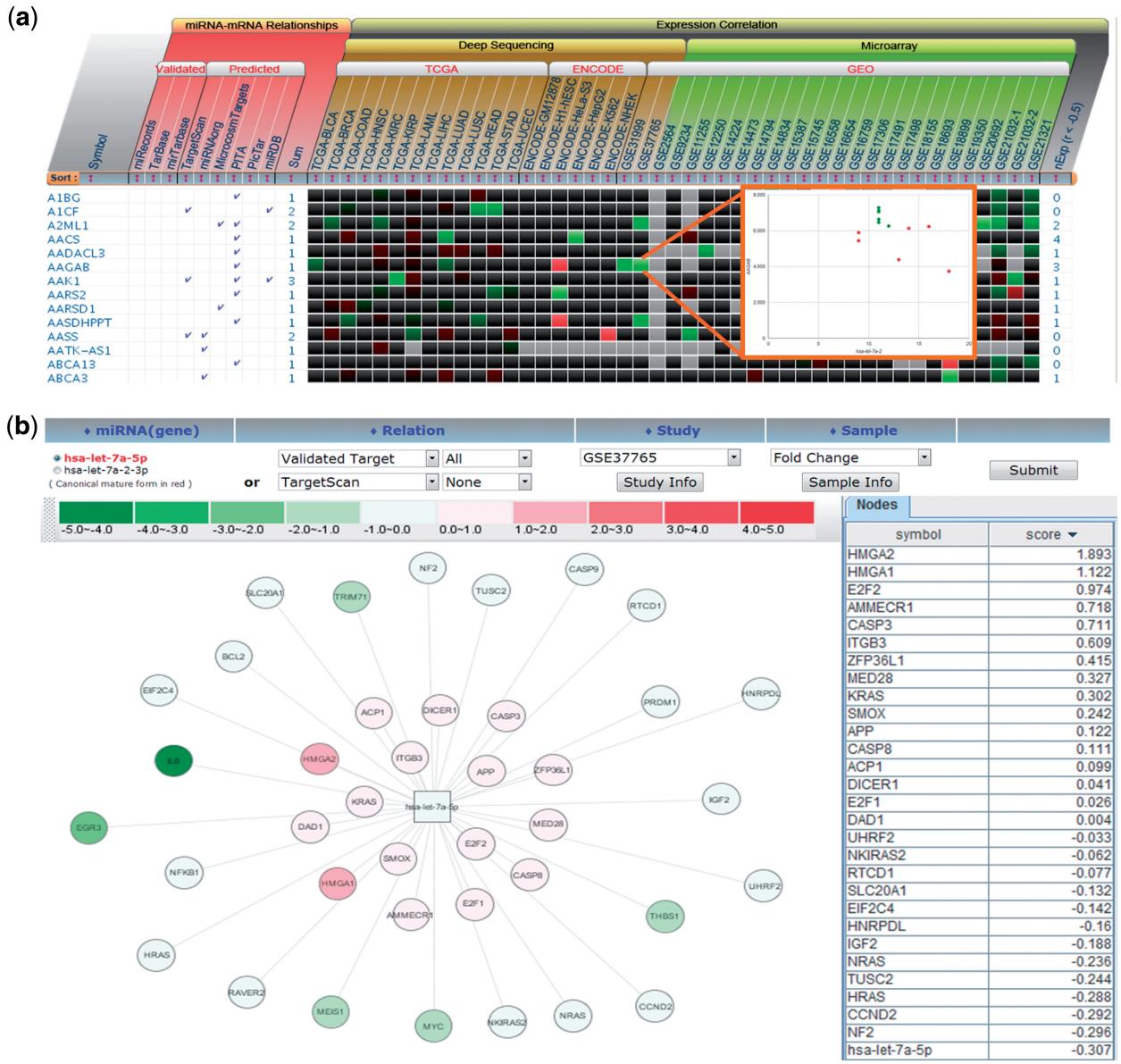


**Figure 3.** Concurrent inspection of miRNA–mRNA target relations and expression correlation for hsa-let-7a-5p. (**a**) The validated and predicted miRNA targets are shown together with their expression correlations as heat-map. The expression values of the miRNA–target pair can be shown for a dataset by clicking a cell as shown in the inset picture. (**b**) An example of miRNA–target network visualization. The targets showing the opposite expression pattern to miRNA are closely placed.

and miRNA–target information is produced with expression correlation as explained in the previous section.

## CONCLUSION

With the addition of deep sequencing data and implementation of several novel tools, miRGator v3.0 continues to be an integrated resource of up-to-date information on miRNA sequences, expression profiling and target identification. These new data and function would be valuable for understanding miRNA biogenesis and molecular functions. However, there are many aspects to improve. Regular update of inundating data is the most critical part since so many sequencing studies are in progress currently including the TCGA project. We plan to update the data annually. Another major advancement in plan is to expand the scope to other organisms such as mice where detailed phenotype information is available via the international mouse phenotyping consortium.

## FUNDING

Korea Research Institute of Bioscience and Biotechnology (KRIBB) Research Initiative Program; National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) [2012-0006002, 2012-0006011,

2012-0366723, 2011-0014992, 2012-0000952]; GIST Systems Biology Infrastructure Establishment Grant (2012) through Ewha Research Center for Systems Biology (ERCSB); Ewha Global Top 5 grant and RP-Grant 2012 from Ewha Womans University. Funding for open access charge: KRIBB Research Initiative Program.

*Conflict of interest statement*. None declared.

## REFERENCES


1. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
2. Zhang,Z., Yu,J., Li,D., Liu,F., Zhou,X., Wang,T., Ling,Y. and Su,Z. (2010) PMRD: plant microRNA database. *Nucleic Acids Res.*, **38**, D806–D813.
3. Szczesniak,M.W., Deorowicz,S., Gapski,J., Kaczynski,L. and Makalowska,I. (2012) miRNEST database: an integrative approach in microRNA search and annotation. *Nucleic Acids Res.*, **40**, D198–D204.
4. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2011) Rfam: wikipedia, clans and the ''decimal'' release. *Nucleic Acids Res.*, **39**, D141–D145.
5. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
6. Vergoulis,T., Vlachos,I.S., Alexiou,P., Georgakilas,G., Maragkakis,M., Reczko,M., Gerangelos,S., Koziris,N., Dalamagas,T. and Hatzigeorgiou,A.G. (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res.*, **40**, D222–D229.
7. Hsu,S.D., Lin,F.M., Wu,W.Y., Liang,C., Huang,W.C., Chan,W.L., Tsai,W.T., Chen,G.Z., Lee,C.J., Chiu,C.M. *et al.* (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **39**, D163–D169.
8. Garcia,D.M., Baek,D., Shin,C., Bell,G.W., Grimson,A. and Bartel,D.P. (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.*, **18**, 1139–1146.
9. Betel,D., Wilson,M., Gabow,A., Marks,D.S. and Sander,C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, **36**, D149–D153.
10. Kertesz,M., Iovino,N., Unnerstall,U., Gaul,U. and Segal,E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
11. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
12. Wang,X. and El Naqa,I.M. (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, **24**, 325–332.
13. Gamazon,E.R., Im,H.K., Duan,S., Lussier,Y.A., Cox,N.J., Dolan,M.E. and Zhang,W. (2010) Exprtarget: an integrative approach to predicting human microRNA targets. *PLoS One*, **5**, e13534.
14. Gennarino,V.A., Sardiello,M., Mutarelli,M., Dharmalingam,G., Maselli,V., Lago,G. and Banfi,S. (2011) HOCTAR database: a unique resource for microRNA target prediction. *Gene*, **480**, 51–58.
15. Liu,H., Jin,T., Liao,R., Wan,L., Xu,B., Zhou,S. and Guan,J. (2012) miRFANs: an integrated database for Arabidopsis thaliana microRNA function annotations. *BMC Plant. Biol.*, **12**, 68.
16. Burroughs,A.M., Ando,Y., de Hoon,M.J., Tomaru,Y., Nishibu,T., Ukekawa,R., Funakoshi,T., Kurokawa,T., Suzuki,H., Hayashizaki,Y. *et al.* (2010) A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome Res.*, **20**, 1398–1410.
17. Yang,J.H., Shao,P., Zhou,H., Chen,Y.Q. and Qu,L.H. (2010) deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res.*, **38**, D123–D130.
18. Zhu,E., Zhao,F., Xu,G., Hou,H., Zhou,L., Li,X., Sun,Z. and Wu,J. (2010) mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res.*, **38**, W392–W397.
19. Hackenberg,M., Sturm,M., Langenberger,D., Falcon-Perez,J.M. and Aransay,A.M. (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.*, **37**, W68–W76.
20. Xie,F., Xiao,P., Chen,D., Xu,L. and Zhang,B. (2012) miRDeepFinder: a miRNA analysis tool for deep sequencing of plant small RNAs. *Plant Mol. Biol.*, **80**, 75–84.
21. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
22. SRA:http://www.ncbi.nlm.nih.gov/sra/ (8 May 2012, date last accessed).
23. TCGA:http://cancergenome.nih.gov/ (13 June 2012, date last accessed).
24. Kelso,J., Visagie,J., Theiler,G., Christoffels,A., Bardien,S., Smedley,D., Otgaar,D., Greyling,G., Jongeneel,C.V., McCarthy,M.I. *et al.* (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.*, **13**, 1222–1230.
25. Nelson,S.J., Schopen,M., Savage,A.G., Schulman,J.L. and Arluk,N. (2004) The MeSH translation maintenance system: structure, interface design, and implementation. *Stud. Health Technol. Inform.*, **107**, 67–69.
26. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
27. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.
28. Kersey,P.J., Staines,D.M., Lawson,D., Kulesha,E., Derwent,P., Humphrey,J.C., Hughes,D.S., Keenan,S., Kerhornou,A., Koscielny,G. *et al.* (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91–D97.
29. Friedlander,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
30. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
31. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
32. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
33. Hofacker,I.L. (2009) RNA secondary structure analysis using the Vienna RNA package. *Curr. Protoc. Bioinformatics*, **Chapter 12**, Unit 12.
34. Griffiths-Jones,S., Crocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
35. Xu,J., Li,C.X., Li,Y.S., Lv,J.Y., Ma,Y., Shao,T.T., Xu,L.D., Wang,Y.Y., Du,L., Zhang,Y.P. *et al.* (2011) MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. *Nucleic Acids Res.*, **39**, 825–836.